Case Study in Geospatial Analytics:
Building a Global Platform for Agro-Environmental Analysis

OGC Agriculture Domain Working Group 2016

09-22-2016 / Kris Matson / Version 2

© Bayer CropScience • September 2016
Bayer CropScience On-Farm Trials
Studying Real-World Product Response

Must understand crop protection product effects in real-world situations

- Efficacy trials on 30 X 30 foot plots do not easily translate to ½ mile X ½ mile fields
  - Not enough variability in the small plot
  - Large fields are not homogeneous
  - Practical conditions differ from R&D

Why answer questions on product response in the real world?

- Help growers to be more profitable – and grow more crops
- Help growers to be more sustainable – limit off-target effects

30ft x 30ft research trial plot

Helping growers to be more profitable and sustainable
Communicating and Managing Trial Protocols

• Originally communicated only at season start
• Insufficient data collection guidance
• Lack of visibility into protocol workflow & issues

Timely Collection of Protocol Data

• Bulk data egress at end-of-season \(\rightarrow\) #epicfail

Missing Metadata

• Metadata is required for on-field activities as well as geospatial data

High-Variety & High-Volume Data

• Equipment, sensors, & FMIS software format all aggregate data differently

Result: Analysis and Modeling At-Scale is Difficult or Even Impossible

• For analytics at-scale, standards are not optional

End of year 1 - complete data from only 50% of trials.
Question: How to scale from 30 to 200 fields?
Answer: STANDARDS

USB is a standard… but what about the files?!
What We Are Building
A Field Trial Protocol Management System → FTPro

Software platform for the systematic collection, processing and analysis of trial data from commercial farm operations

Provides standardized data to visualization and analytic tools:

• To understand product performance and ROI
• Analyze multiple fields and seasons
Inspiration for this Presentation
Vision for Geospatial Analytics via Open Standards

The Future of Geospatial Analytics through Open Standards

George Percivall
CTO, Chief Engineer
Open Geospatial Consortium
percvall@myogc.org

© 2016 Open Geospatial Consortium
OGC Future of Geo-Analytics
Lack of interoperability is a serious technical debt

Lessons from the success of Apache Spark...
interchange is necessary for the ecosystem
major use cases tend to build their own ML libraries – despite a case
where a majority of committers tend to support a common vision and
encourage use of a canonical library (MLLib with DataFrames)
when a successful business grows over time, challenges arise by
definition: managing separated teams, mergers and acquisitions,
increased audits, regulations, etc.

therefore, lack of interchange for analytics represents a serious
technical debt and potential liability

Source: “Use of standards and related issues in predictive analytics” KDD 2016,
SF 2016-08-16  Paco Nathan, O’Reilly Media
OGC Big Geo Data Analysis Use Case
Presented at ENVI Advanced Analytics Symposium

Big Geo Data

- High Velocity Ingest
- Geospatial Databases
- GeoAnalytics, Machine Learning
- Spatial Modeling

Observation Sources

Users and consuming apps

© 2016 Open Geospatial Consortium
## Big Data Use Case for Ag R&D Trials
### High-Variety & High-Volume Analytic Pipeline

<table>
<thead>
<tr>
<th>High Variety &amp; Volume Ingest</th>
<th>Geospatial Databases</th>
<th>Data Aggregation and Analysis</th>
<th>Visualization &amp; Modeling</th>
</tr>
</thead>
<tbody>
<tr>
<td>Agronomic &amp; Management Data</td>
<td>Relational databases</td>
<td>Data QA - Cleanse - Impute</td>
<td>Visualization &amp; Workflow Systems</td>
</tr>
<tr>
<td>IoT Sensor Data</td>
<td>No-SQL databases</td>
<td>Data Fusion</td>
<td>Component Agro-Enviro models</td>
</tr>
<tr>
<td>Laboratory Data</td>
<td>Cloud APIs</td>
<td>Entity-oriented Spatial-temporal analytics</td>
<td>Integrated Agro-Environment Simulations</td>
</tr>
<tr>
<td>Proprietary Data Sources &amp; APIs</td>
<td>Unstructured Data</td>
<td>Grid-oriented Spatial-temporal analytics</td>
<td>Statistical Processing Systems</td>
</tr>
<tr>
<td>Open Source Data &amp; APIs</td>
<td></td>
<td>Remote sensed data processing</td>
<td>Machine Learning Systems</td>
</tr>
</tbody>
</table>

**Business Value**
# Typical Field Data Sources

Agronomic, Management, and Spatial data

<table>
<thead>
<tr>
<th>Agronomist + Grower Collected Data</th>
<th>IoT Equipment Generated Data</th>
</tr>
</thead>
<tbody>
<tr>
<td>- Field Boundary</td>
<td>- Planting (Seeding)</td>
</tr>
<tr>
<td>- Field Scouting</td>
<td>- As-Applied Fertilizer</td>
</tr>
<tr>
<td>- Soil Cores</td>
<td>- As-Applied Herbicide</td>
</tr>
<tr>
<td>- Soil Chemical Analysis</td>
<td>- As-Applied Pesticide</td>
</tr>
<tr>
<td>- Crop Tissue Samples</td>
<td>- As-Applied Insecticide</td>
</tr>
<tr>
<td>- ...</td>
<td>- Harvest (Yield)</td>
</tr>
<tr>
<td></td>
<td>- Weather Stations</td>
</tr>
<tr>
<td></td>
<td>- ...</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Public Sector Data</th>
<th>Multi-Source Imagery</th>
</tr>
</thead>
<tbody>
<tr>
<td>- Elevation</td>
<td>- UAS, Air-borne, Satellite</td>
</tr>
<tr>
<td>- Soil</td>
<td>- NDVI and other derivative products</td>
</tr>
<tr>
<td>- Landsat</td>
<td></td>
</tr>
<tr>
<td>- ...</td>
<td></td>
</tr>
</tbody>
</table>
Agriculture IoT Data
Equipment data capture and aggregation

Typical Ag equipment data issues:
• Was the seed variety entered into the Planter?
• Was the applied product entered?
• Was an applicator nozzle clogged?
• Was the harvester calibrated?
• Data egress in cloud or as files?
Agronomist & Grower Collected Data
Varying sources, structure, aggregation and standards

Field scouting

Farm management information system

Cloud

Field digitizing

Management recommendations

shp, csv, ...

Report documents
Sensors and Multi-Source Imagery
Covariate data capture and aggregation

**Unmanned Air Systems**
- Airborne imagery
- Satellite imagery

**Weather and Moisture**
- Vendor APIs
- Bayer CS Aggregation Services

Vendor APIs
SST agX Cloud
Standards and AWS interface for field data

- Standardized cloud interface for agriculture operations data
  - Field-specific data payloads
  - Management data
  - Spatial data
- XML encoding
  - Robust schemas XSD encoded
  - WKT shape types for geometries
  - GeoTIFF for imagery
- Practical standard supporting:
  - Farm operations
  - System interoperability

Farm management information system
SST agX Cloud
Soil sampling XSD

```xml
<xs:schema targetNamespace="http://www.stsoft.com/EDS/SoilSamplingFullTask.xsd" elementFormDefault="qualified" xml:space="preserve">
  <xs:element name="SoilSampleFull">
    <xs:complexType>
      <xs:sequence>
        <xs:element name="Record" maxOccurs="unbounded">
          <xs:complexType>
            <xs:sequence>
              <xs:element name="RecordNum" type="xs:unsignedInt"/>
              <xs:element name="ZoneID" minOccurs="0"/>
              <xs:element name="SampleID"/>  
              <xs:element name="TopsoilSamplingDepth" minOccurs="0"/>
              <xs:element name="Soil_pH" minOccurs="0"/>
              <xs:element name="Buffer_pH" minOccurs="0"/>
              <xs:element name="ExcessLime" minOccurs="0"/>
              <xs:element name="PercentCaCO3" minOccurs="0"/>
              <xs:element name="H_ineq" minOccurs="0"/>
              <xs:element name="ExchangeAcidity" minOccurs="0"/>
              <xs:element name="PercentOC" minOccurs="0"/>
              <xs:element name="PercentHumicMatter" minOccurs="0"/>
              <xs:element name="TotalN" minOccurs="0"/>
              <xs:element name="NO3_N" minOccurs="0"/>
            </xs:sequence>
          </xs:complexType>
        </xs:element>
      </xs:sequence>
    </xs:complexType>
  </xs:element>
</xs:schema>
```
Soil Sampling Task v 2.0

Bayer to evaluate this next quarter – global objective

<?xml version="1.0" encoding="UTF-8"?>
  <SyncID>20</SyncID>
  <ParentEventID>dcf0831c-4a91-41bb-9a77-6ae2367e2e71</ParentEventID>
  <ModifiedOn>2016-07-14T16:19:33.248Z</ModifiedOn>
  <Records>
    <Record>
      <RecordNum>0</RecordNum>
      <SampleID>502</SampleID>
      <TopsoilSamplingDepth>
        <Depth>12</Depth>
        <Unit>504</Unit>
        <ID>52</ID>
        <Name>in</Name>
      </TopsoilSamplingDepth>
      <Soil_pH>6.9</Soil_pH>
      <OM>
        <Measure>3</Measure>
        <Unit>1217</Unit>
        <ID>329</ID>
        <Name>Percent</Name>
      </OM>
    </Record>
    <Record>
      <RecordNum>1</RecordNum>
      <SampleID>502</SampleID>
      <Phosphorus>
        <ExtractionMethod>Bray 1</ExtractionMethod>
        <Measure>70</Measure>
        <Unit>740</Unit>
        <ID>324</ID>
        <Name>ppm</Name>
      </Phosphorus>
    </Record>
  </Records>
</SoilSample>
FTPro Architecture
High-Level System Object Model

- Visualization & Protocol Management
  - Protocol Management
  - Reporting
  - Mapping
  - EDA + Data Metrics

- Equipment Sensors
  - Multi-Source
  - High-Variety

- Human Observation
  - Scouting
  - Management Info
  - Process dependent

- Data Aggregation Services
  - Data QA & Cleanse
  - Interpolation
  - Gridding

- Database Management System
  - Unstructured – L1
  - RDBMS – L2
  - Analytic Store – L3

- Data Source Adapters
  - Data Sources Ingest
  - Multiple Standards
  - Schema Mapping

- Analytics & Modeling
  - Geo Analytics
  - Statistics
  - Machine Learning
  - Agro-Enviro Models

- Non-Ag Sensors
  - WX
  - Imagery

- Open Data Services
  - Soils
  - Elevation

- S3 Unstructured
- PostGIS Shapes
- Postgres Structured
- Loose File
- Cloud APIs
- Various Stds

- Loos File
- Cloud APIs
- Various Stds
Principles of our Approach

Iterations on a Minimum Viable Product

- **Standardize the workflow UI and UX in software first**
  - Maximizes Product Owner participation at the start
  - *Then* iteratively re-engineer the back-end and interfaces as demanded by user stories

- **Maximize use of Open Systems and Open Standards in early platform development phases**
  - Core back-end systems are the hardest to change
  - Use COTS proprietary systems for fast build-out where expedient

- **Maximize use of loosely coupled web services for platform interfaces**
  - It is easier to use other’s interfaces than design and build your own
  - De-couple interfaces later where proprietary interfaces or tight-coupling was used in early iterations, guided by technical debt and/or user stories

- **Be highly aware of accumulating technical debt**
  - Keep track in the product backlog
  - Standards are practical and essential but require thinking beyond the next 2 sprints
Standardizing the Protocol Contract
Template the protocol and its component parts

Each protocol and sub-elements are configurable templates

- The Job is a template
- The Job Step is a template

Allows flexibility for each experiment while governing inputs
Standardizing the Protocol Workflow
Adding Context – Geospatial Data is not Enough

Contextual data varies depending on the job step

- Provides a flexible way to collect field and protocol management data without making changes to the data model
- The seed variety may be very important for understanding the harvest
- Equipment manufacturer will be important for application, but not for soil sampling
Standardizing the Analytic Pipeline

JSON Templates for Analytics

Each data set type has unique spatial processing sequences

- Spatial processing templates are easily configured and stored as JSON objects
- Gives users the ability to tailor analytics to the research objectives and source of data collection

Syntax

EmpiricalBayesianKriging_ga (in_features, z_field, {out_ga_layer}, {out_raster}, {cell_size}, {transformation_type}, {max_local_points}, {overlap_factor}, {number_semivariograms}, {search_neighborhood}, {output_type}, {quantile_value}, {threshold_type}, {probability_threshold}, {semivariogram_model_type})
Each internal geospatial data type has a standard schema

- Interactive schema mapping transforms data from any source to a common Level 2 data model for visualization, plus first-order analytics and modeling
- Deeper analytics go to the Level 1 data store
- Visualization symbology is also a user-modifiable template
Weather Service for On-Field Stations

Private weather station aggregation architecture v.1

- Each station xmits raw to Davis weatherlink cloud, 15 min intervals
- Weatherlink aggregates, publishes JSON messages
- BCS AWS service subscribes and aggregates JSON in RDBMS
- BCS publishes REST interfaces for consuming apps
Thank you!
Kris Matson -- Geo-Analytics, Data Visualization, Software Development

Kris.Matson@bayer.com / kmatson@lifescaleanalytics.com / m: 919-810-1839